Web 主题关联知识自学习算法*)

杨 沛 郑启伦 彭 宏
(华南理工大学计算机科学与工程学院 广州 510640)

Discovery of Web Topic-Specific Association Rules

YANG Pei ZHENG Qi-Lun PENG Hong
(College of Computer Science and Engineering, South China University of Technology, Guangzhou 510640)

Abstract There are hidden and rich information for data mining in the topology of topic-specific websites. A new topic-specific association rules mining algorithm is proposed to further the research on this area. The key idea is to analyze the frequent hyperlinked relations between pages of different topics. In the topic-specific area, if pages of one topic are frequently hyperlinked by pages of another topic, we consider the two topics are relevant. Also, if pages of two different topics are frequently hyperlinked together by pages of the other topic, we consider the two topics are relevant. The initial experiments show that this algorithm performs quite well while guiding the topic-specific crawling agent and it can be applied to the further discovery and mining on the topic-specific website.

Keywords Association rule, Topic-specific crawling, Web mining

1 概述

面向主题的 Web 信息搜索和挖掘是当前的一个研究热点,它在一定的应用背景下取得了很大的成功,如 Cora[1]和 CiteSeer,但与此同时也存在很多尚待解决的问题,其中包括:第一,网页搜索没有能够充分利用搜索过程中网页与网页、网页与链接,以及链接与链接之间相互关联与约束的有关知识,因而无法更有效地提高搜索的效率和搜索的准确性。第二,网页搜索知识的获取需要用户提供,或者事先提供大量具有代表性的训练样本自学获得,因此网页搜索知识有效获取问题依然存在。

引入关联规则挖掘[2,3]等技术是解决这些问题的一个有效途径。但是,目前的 Web 关联规则挖掘一般是针对 Web 使用日志[4,5],通过对用户访问系列进行挖掘,从中发现有趣的模式。而对领域型网站内部拓扑结构和主题组织意图进行挖掘的研究并不多。事实上,领域型网站的拓扑结构中蕴含了大量人类潜在的语义和主题关联知识,这都是可供挖掘的重要资源。文[6]利用"共引用"的思想来挖掘相关主题,文[7]使用了扩展的概念分层技术来对半结构化数据进行关联规则挖掘。这些方法的一个主要缺点是没有能够充分挖掘领域型网站中各种不同类型和不同层次的主题关联关系。

为此,本文提出了一种新的面向主题的关联规则挖掘和自学习方法。其主要思想是:在领域型网站中,如果两个不同主题的网页集之间存在频繁的链接关系,则认为这两个主题是相关的;如果两个不同主题的网页集与另一主题的网页集之间均存在频繁的链接关系,则认为这两个主题是相关的。主题关联规则在 Web 主题信息的搜索和挖掘领域有着很重要的应用,如:用主题关联规则解决网页搜索中的最优搜索路径的选择问题;用主题关联规则对网页进行分类;解决主题搜索和挖掘中的知识表示、知识推理和知识获取问题。

2 网络主题关联模型

设领域词典表示为 $T=\{t_i|t_i$ 为特征项,$i=1,2,\cdots,n\}$

定义 1(网页特征集,CT) 网页特征集是指网页中或是指向网页的链接中包含的特征项的集合。网页特征集 CT 的构造方法如下,设指向网页的链接为 l,网页的标题为 title,对链接文本和网页标题进行词条切分和停用词处理后得到特征项集合 T:

$$CT(v)=\begin{cases}\{t_i|t_i \text{ 在链接 } l \text{ 的文本中}, i\in N\} & v \text{ 不是首页}\\ \{t_i|t_i \text{ 在网页标题中}, i\in N\} & \text{否则}\end{cases}$$

之所以用指向网页的链接信息来构造 CT,原因有两个:首先,网页本身缺乏自描述信息,而且包含过多的噪声数据;第二,指向网页的链接中包含了网页引用者对该网页更加准确精炼的解释。

定义 2(网页节点) 网页节点用一个四元组表示,即 v=(DocId, CT, R, timestamp),其中,DocId 唯一标识一个网页,CT 为网页特征项集,R 为网页的主题相关度,通过余弦法计算得到,timestamp 为获取网页的时间。

这样,对于领域型网站,可以用拓扑图 G=(V, E)表示,其中,V 是网页的集合,E 是网页之间的超链接集合。V 中每个节点用 v=(DocId, CT, R, timestamp)表示。

定义 3(特征项节点集) 特征项节点集是指包含该特征项的所有网页节点的集合,记为:$V(t_i)=\{v_j|v_j\in V, t_i\in CT(v_j)\}$,其中 $CT(v_j)$表示节点 v_j 的网页特征集。$|V(t_i)|$表示集合中网页节点元素的个数。

定义 4(节点收益) 一个网页节点的收益是指该网页节点的所有后继节点的主题相关度按一定相关系数打折后的线形和。记为:

$$I(v)=\sum_{u\in F_v}\gamma\cdot R(u) \qquad (1)$$

其中,R(u)为节点的相关度,F_v 表示节点 v 的后继节点集。

*)本课题得到国家自然科学基金(60003019),广东省自然科学基金(990582),广东省科技攻关项目(C10201)资助。杨 沛 博士生,主要研究方向:Web 主题搜索、Web 挖掘、人工智能。郑启伦 教授,博士生导师,主要研究方向:人工智能等。

3 主题关联规则及兴趣度度量

3.1 主题关联规则

定义 5(主题关联规则 Rule) 特征项之间的主题关联关系称为主题关联规则，记为：$Rule(t_i \Rightarrow t_j)$，其中 t_i、t_j 为特征项。

定义 6(主题关联知识库 RulesSet) 所有主题关联规则的集合就构成主题关联知识库。记为：

$RulesSet = \{Rule(t_i \Rightarrow t_j) \mid Rule(t_i \Rightarrow t_j)$为主题关联规则$\}$

在领域型网站中，存在不同类型和不同层次的主题关联关系，下面分三种情况讨论。如图 1 所示，t_i、t_j、t_k 为特征项，圆圈表示特征项所在的节点，实线表示节点之间的链接关系，虚线表示特征项之间的主题关联关系。

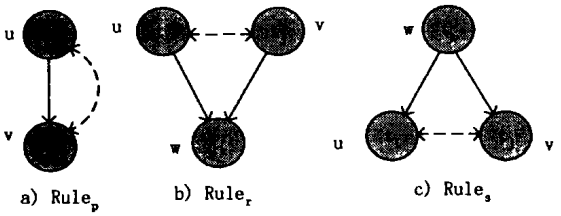

图 1 不同的主题相关关系

(a)设 u、v 为网页节点，t_i、t_j 为特征项，若 $t_i \in CT(u)$，$t_j \in CT(v)$，且 $u \rightarrow v$，则可以得到主题关联规则：

$$Rule_p(t_i \Rightarrow t_j) \tag{2}$$

(b)设 u、v、w 为网页节点，t_i、t_j、t_k 为特征项，若 $t_i \in CT(u)$，$t_j \in CT(v)$，$t_k \in CT(w)$，且 $u \rightarrow w$，$v \rightarrow w$，则可以得到主题关联规则：

$$Rule_r(t_i \Rightarrow t_j) \tag{3}$$

(c)设 u、v、w 为网页节点，t_i、t_j、t_k 为特征项，若 $t_i \in CT(u)$，$t_j \in CT(v)$，$t_k \in CT(w)$，且 $w \rightarrow u$，$w \rightarrow v$，则可以得到主题关联规则：

$$Rule_s(t_i \Rightarrow t_j) \tag{4}$$

3.2 兴趣度度量

在主题关联规则的挖掘过程中，需要限制不感兴趣的模式数量，这可以通过设定兴趣度度量来实现。兴趣度度量评估模式的简洁性、确定性、实用性和新颖性。在本系统中，除了采用常规的置信度和支持度来裁剪兴趣度较低的模式之外，还采用主题度来评估模式的主题搜索性能。

(1)置信度

$$Confidence(Rule(t_i \Rightarrow t_j)) = p(t_j \mid t_i) = \frac{p(t_i, t_j)}{p(t_i)} \tag{5}$$

(2)支持度

$$Support(Rule(t_i \Rightarrow t_j)) = p(t_i, t_j) \tag{6}$$

(3)主题度

$$Topic(Rule(t_i \Rightarrow t_j)) = \frac{\sum_{v \in V(t_j)} I(v)}{|V(t_j)|} \tag{7}$$

其中，$V(t_j)$为 t_j 的特征项节点集，$I(v)$为 v 的节点收益，上式可解释为：将 t_j 的特征项节点集中所有节点的主题相关度的数学期望值作为规则 $Rule(t_i \Rightarrow t_j)$的主题度。主题度越高的模式就越可能引导搜索 agent 找到主题相关网页。

主题关联知识库中的主题关联规则应满足：

$$\sum_{Rule(t_i \Rightarrow t_j) \in RulesSet} Topic(Rule(t_i \Rightarrow t_j)) = 1 \tag{8}$$

对于主题关联规则 $Rule(t_i \Rightarrow t_j)$，若满足最小置信度、最小支持度和最小主题度，则称之为强主题关联规则。

4 主题关联知识挖掘和自学习

面向主题的知识表示、知识获取和知识推理是 Web 主题信息搜索和挖掘面临的一个关键问题。为了有效地解决这个问题，本文提出了一种新的主题关联知识挖掘和自学习方法。

主题关联知识挖掘和自学习主要分三个阶段：第一阶段是训练阶段，搜索 agent 选取若干个典型的领域型网站作为训练集，对这些网站进行搜索，从中提取主题关联规则。第二阶段是测试阶段，搜索 agent 使用主题关联规则引导主题搜索。第三阶段是反馈和自学习阶段，搜索 agent 根据搜索结果更新主题关联知识库。

4.1 主题关联规则挖掘

主题关联规则挖掘的主要过程：从 YAHOO 分类目录中选取若干个典型的领域型网站作为训练集。搜索 agent 对训练集中包括的网站展开搜索，并根据余弦法计算网页的主题相关度，根据一定的阈值剔除噪声网页。然后，利用算法 1 对搜索到的网页进行挖掘，提取主题关联规则。

算法 1 主题关联规则挖掘

(1)设 $P = \{p_1, p_2, \cdots, p_n\}$为搜索到的所有网页的集合。根据余弦法计算网页的主题相关度。

(2)根据网页集 P 中网页之间的链接关系构造网络主题关联拓扑图 G。根据式(1)计算节点收益。

(3)从图 G 中入度为零的节点开始，逐个处理图 G 中每个节点。

(4)设当前处理节点为 u，根据图 G 得到 u 的所有后继节点集 $F_u = \{v_i \mid v_i$ 为 u 的后继节点，$i \in N\}$。逐个处理每个节点对(u，v)：

节点 u 及其子节点 v 的特征项集作连接运算，得到 2-项集，记为 CL_2：

$$CL_2 = CT(u) \infty (CT(v) = \{(t_{ui}, t_{vj}) \mid t_{ui} \in CT(u), t_{vj} \in CT(v), t_{ui} \cap t_{vj} = \Phi\}$$

(5)图 G 中所有的节点都处理完后，得到一个 2-项集的集合。在 2-项集的集合中挖掘频繁 2-项集，记为 L_2。

(6)根据式(2)、(3)、(4)，在频繁 2-项集 L_2 中寻找候选主题关联规则。

(7)根据式(7)计算候选主题关联规则的主题度，并根据式(8)对关联规则主题度进行范化。

(8)根据式(5)、(6)、(7)从候选主题关联规则中寻找强主题关联规则。

4.2 主题关联规则引导主题搜索

利用主题关联规则引导主题搜索[8]的主要思想是，通过主题关联规则预测待扩展的节点的权值，搜索 agent 优先搜索权值较大的节点。

算法 2 用主题关联规则引导搜索(ARS)

(1)初始化，URL 种子集添加到 URL 队列。

(2)如果 URL 队列为空，则搜索结束。

(3)取出 URL 队列头节点，根据 URL 获取网页。

(4)对获取到的网页节点进行处理。

设当前处理节点为 u，找到 u 的所有后继节点集 $F_u = \{v_i \mid v_i$ 为 u 的后继节点，$i \in N\}$。逐个处理节点对(u，v)：

两个节点的特征项集作连结运算，得到 2-项集：

$CL_2 = CT(u) \infty CT(v) = \{(t_{ui}, t_{vj}) \mid t_{ui} \in CT(u), t_{vj} \in CT$

$(v), t_{wi} \cap t_{vj} = \Phi\}$

根据(t_{wi}, t_{vj})检索主题关系知识库，得到主题关联规则 $Rule(t_{wi} \Rightarrow t_{vj})$的主题度，取 CL_2 中主题度最大者作为节点 v 的预测权重(如果没命中，则置为零)，即：

$$R'(v) = \text{argmax}_{(t_{wi}, t_{vj}) \in CL_2}(T(Rule(t_{wi} \Rightarrow t_{vj})))$$

并将节点 v 放入 URL 队列。

(5)根据节点的预测权重对 URL 队列进行排序。转(2)。

4.3 主题关联知识自学习算法

主题关联知识自学习算法的主要思想是：对每次搜索结果进行主题关联规则挖掘，并更新原来的主题关联知识库，从而自动地对主题关联知识库进行增加、修改、删除，实现主题关联知识自学习。

算法 3 主题关联知识自学习算法

(1)调用主题关联规则挖掘算法对搜索结果进行处理，得到新的临时主题关联知识库，记为 RulesSet2。

(2)根据 RuleSets2 更新原有的主题关联知识库 RulesSet。分三种情况：

a) 对于 RulesSet2 中的主题关联规则 $Rule2(t_i \Rightarrow t_j)$ $(Rule2 \in RulesSet2)$，如果在 RulesSet 中存在对应的主题关联规则 $Rule(t_i \Rightarrow t_j)(Rule \in RulesSet)$，则取两者的兴趣度的平均值作为新的兴趣度。即有：

$$C'(Rule(t_i \Rightarrow t_j)) = \frac{C(Rule(t_i \Rightarrow t_j)) + C(Rule2(t_i \Rightarrow t_j))}{2}$$

$$S'(Rule(t_i \Rightarrow t_j)) = \frac{S(Rule(t_i \Rightarrow t_j)) + S(Rule2(t_i \Rightarrow t_j))}{2}$$

$$T'(Rule(t_i \Rightarrow t_j)) = \frac{T(Rule(t_i \Rightarrow t_j)) + T(Rule2(t_i \Rightarrow t_j))}{2}$$

b) 对于 RulesSet2 中的主题关联规则 $Rule2(t_i \Rightarrow t_j)$ $(Rule2 \in RulesSet2)$，如果在 RulesSet 中不存在对应的主题关联规则 $Rule(t_i \Rightarrow t_j)(Rule \in RulesSet)$，则将 $Rule2(t_i \Rightarrow t_j)$添加到主题关联知识库 RulesSet 中。

c) 对于 RulesSet 中的主题关联规则 $Rule(t_i \Rightarrow t_j)(Rule \in RulesSet)$，如果在 RulesSet2 中找不到对应的规则，则降低该规则的主题度，即有：

$T'(Rule(t_i \Rightarrow t_j)) = \gamma \cdot T(Rule(t_i \Rightarrow t_j))$，其中 γ 为衰减系数，$0 \leqslant \gamma \leqslant 1$

5 实验分析

在我们开发的电子商务智能搜索和挖掘系统(Ego)上对主题关联规则引导搜索算法(ARS)和 BreadthFirst 算法进行了对比测试。Ego 系统目前主要是搜索移动科技产品方面的网站。系统采用分布式体系结构，多台机器同时搜索。实验结果如图 2 所示，x 轴表示已经搜索过的网页数，y 轴表示主题保持度。主题保持度是指所搜索网页的主题相关度的数学期望。其中，网页的主题相关度可通过余弦法计算得到。

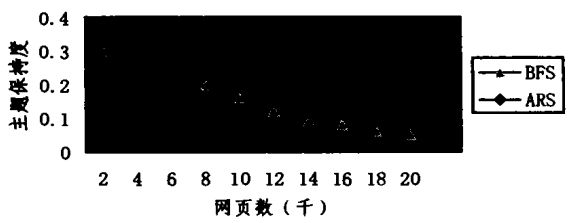

图 2 搜索网页及其平均相关度

从图中可以看出，在搜索过程中，ARS 保持了较高的主题保持度。BFS 在搜索初期尚能保持较高的平均主题度，随着搜索范围的扩大，BFS 的主题保持度下降较快。而 ARS 方法则较为平稳。同时，对搜索到的网页进行分析可以发现，在搜索的网页数超过一万个以后，基于 BFS 算法的索引库中包含了大量的与移动科技产品无关的噪声网页。也就是说，在远离搜索种子集之后，BFS 很快就发生了主题漂移。基于 ARS 算法的搜索 agent 也会搜索到部分的与主题无关的网页，但是它能很快地根据主题关联规则进行自调整，将主题无关的 URL 裁剪掉，从而迅速地跳出搜索误区，使搜索能保持在与主题相关的区域内。

结语 Web 主题搜索和挖掘是一个极具挑战性的研究领域。而面向主题的知识表示、知识推理和知识获取则是要首先解决的一个关键问题。本文提出的 Web 主题关联知识挖掘和自学习算法在这方面作了有益的尝试。与其它方法相比，该方法的一个主要优点是：它不需要预先提供大量的训练样本，而是着重于挖掘领域对象内在蕴含的大量主题相关知识，从中自动地获取知识，并将这些知识应用到主题搜索和挖掘过程中；与此同时，它还利用反馈来实现主题知识的自学习。

下一步的研究重点是，在本文研究的基础上，进行多维多层的主题关联知识的挖掘，从而构建一个“知识网”，并利用这个“知识网”进行主题搜索和挖掘、网页分类和网页聚类等，即所谓的“以网对网”的思想。

参考文献

1 McCallum A, Nigam K, Rennie J, et al. Building domain-specific search engines with machine learning techniques. In: Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999

2 Lee C H, Yang H C. Developing an adaptive search engine for e-commerce using a web mining approach. In: Proc. of the Intl. Conf. on Information Technology: Coding and Computing, 2001. 604～608

3 Nahm U Y, Mooney R J. Text mining with information extraction. In: AAAI2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, Standford, CA, 2002

4 Mobasher B, Dai H, Luo T, et al. Effective personalization based on association rule discovery from web usage data. Web Information and Data Management. 2001

5 Clifton C, Cooley R, Rennie J. TopCat: data mining for topic identification in a text corpus. Principles of Data Mining and Knowledge Discovery. 2002

6 Yi J, Sundaresan N. Metadata based web mining for relevance. 2000 (IEEE) International Database Engineering and Applications Symposium (IDEAS'00)

7 Singh L, Scheuermann P, Chen B. Generating association rules from semi-structured using an extended concept hierarchy. web. ece. nwu. edu/EXTERNAL/dbwww/papers/CIKM97. ps.

8 Glover E J, Tsioutsiouliklis K, Lawrence S, et al. Using web structure for classifying and describing web pages. In: WWW2002, Honolulu, Hawaii, USA